单样本率比较（单组目标值法）样本量计算不同方法的比较

曾治宇 李青
波科国际医疗贸易（上海）有限公司

【摘要】 目的 对单样本率比较（单组目标值法）样本量计算的不同方法进行比较，为实际应用中选择合适的方法提供依据。 **方法** 构建目标值π_0和预计值π_1，以正态近似法、通用法、反正弦法、确切经典法及确切保守法等 5 种方法分别计算各自所需的样本量，编程计算相应的最低成功率，并进行计算机模拟获得检验效能。 **结果** 5 种方法在π_0及π_1不接近 0 或 1 时表现较为相似，但π_0逐渐接近 0 时，正态近似法和通用法得到的样本量相对较小，并逐渐损失了检验效能；π_0逐渐接近 1 时，正态近似法和通用法得到的样本量相对较大，检验效能也比预设值逐渐增高。从检验效能来看，反正弦法的结果与确切经典法接近而显得更为离散，而确切保守法几乎能保证预设的检验效能，但在π_0>0.5 时，确切保守所需样本量比确切经典法逐渐增加。不同方法对实际成功率的要求总体相似，但存在细小差别。 **结论** 单个率比较的样本量计算方法的选择较为复杂，对检验效能要求比较高时，宜优选确切经典法和确切保守法，其次可考虑反正弦法，而通用法和正态近似法在率偏向两侧时，样本量会过大或过小，应具体权衡。

【关键词】 单样本率比较 单组目标值 样本量计算

单组目标值法的临床研究设计近年来受到越来越多的关注，特别是在医疗器械的临床试验中[1]。计量资料的单组目标值法的样本量计算方法较为一致，而对于计数资料，样本量计算的方法仍有不同的考虑。

结果为二分类的计数资料的单组目标值法的本质是单样本率与已知总体率的比较，事先确定一个目标值π_0（总体率），设定显著性水平 α（通常为单侧检验，取值 0.025）和检验效能 1-β（β 通常取值 0.2 或 0.1），然后将本研究组预计达到的结果π_1（样本率）与之比较，从而获得样本量。假设检验时，如果是高优指标（如治疗成功率），当π_1的 1-2α 可信区间的下限不小于π_0时则拒绝无效假设，接受$\pi_1 \geqslant \pi_0$，低优指标（如并发症）同理。

目前单样本率比较的样本量计算通常有以下 5 种方法。一些经典的统计学教科书给出的基于正态近似法的样本量计算公式为（以下称为正态近似法）[2]:

$$n = \pi_0(1-\pi_0)[\frac{(Z_\alpha + Z_\beta)}{(\pi_1 - \pi_0)}]^2$$

其他一些统计学教科书[3]、中国临床试验生物统计学组[4]及国家药品监督管理局[5]等推荐的公式为（以下称为通用法）:

$$n = \frac{[Z_\alpha\sqrt{\pi_0(1-\pi_0)} + Z_\beta\sqrt{\pi_1(1-\pi_1)}]^2}{(\pi_1 - \pi_0)^2}$$

当π_0或π_1接近 0 或 1 时正态性较差，宜考虑数据转换，根据平方根反正弦转换的样本量计算公式为（以下称为反正弦法）[3]:

$$n=\frac{(Z_\alpha+Z_\beta)^2}{4(sin^{-1}\sqrt{\pi_1}-sin^{-1}\sqrt{\pi_0})^2}$$

或者使用基于二项分布理论的确切概率法计算样本量[6]。由于单样本率确切概率法获得的样本量与检验效能非单调递增[7, 8]，会有一个常规的结果和一个较为保守的结果，本文称之为确切经典法和确切保守法。

这 5 种样本量计算的方法孰优孰劣，具体的适用条件如何，尚没有系统的研究。本文的目的在于利用计算机模拟分析，考察不同方法计算的样本量及其相应的实际成功率及检验效能，探索不同方法的使用条件，为实际应用中选择合适的方法提供依据。

方法

1 构建π_0和π_1

为方便起见，仅考察高优指标。构建π_0从 0.01 至 0.98，按 0.01 递增。不失一般性，当π_0在 0.01，[0.02, 0.04], [0.05, 0.19], [0.2, 0.79], [0.8, 0.94], [0.95, 0.97]及 0.98 时，π_1分别增加 0.01，0.02，0.05，0.1，0.05，0.02 及 0.01。低优指标转可根据率的对称性化为高优指标，比如某研究的并发症的目标值为 10%，预计研究组的并发症可降至 5%，可将低优指标并发症转化为高优指标成功率，即成功率的目标值为 90%，预计研究组可提高至 95%。

2 不同样本量计算方法的评价

2.1 样本量计算

基于构建的π_0和π_1，设定 α=0.025（单侧），β=0.2，根据上面介绍的正态近似法、通用法、反正弦法、确切经典法及确切保守法 5 种方法，分别计算各自所需的样本量。

2.2 实际成功率

根据计算出的样本量，编程计算求得所需的最小成功例数，即可获得实际的成功率。所需的最小成功例数为满足该成功率 95%可信区间下限≥π_0所需的最小例数。95%可信区间的构建需与样本量计算的方法一致，即正态近似法及通用法采用正态近似的方法构建，反正弦法采用平方根反正弦转换后再进行正态近似的方法构建，而确切经典法和确切保守法根据二项分布的理论构建（Clopper-Pearson 可信区间）。

2.3 检验效能

采用计算机模拟计算的方法获得检验效能。根据已知样本量及预计成功率π_1，进行二项分布概率抽样获得成功例数，重复 10000 次，统计成功例数≥所需最小成功例数的次数，除以重复次数即为检验效能。

本研究的评价指标包括样本量，实际成功率和检验效能。一般来说，在满足 α 及 β 的情况下，样本量越小越好，检验效能则越高越好。而对于实际成功率，监管部门通常希望看到更高的成功率，而研究者及厂家为获得阳性结果，通常更希望能够以较低的成功率达到假设检验的显著性。

3 实例分析

实例 1 来自《人工耳蜗植入系统临床试验指导原则》[9]：根据临床经验，开机 12 个月后，产品的总体有效率需至少达到 70%（目标值为 70%）方可被临床接受。假设被试验产品的总体有效率可以达到 85%，计算在双侧显著性水平 0.05、把握度 80%的情况下的样本量。实例 2 来自文献[7]，已知总体率π_0=0.07，预期的总体率π_1=0.03，设定 α=0.05（双侧），检验效能 80%，计算样本量。该例子可验证低优指标的情况。实例 3 设定π_0=0.1，π_1=0.2，可验证一下π_0<0.5 的情形。实例 4 来自一个真实的全皮下 ICD（S-ICD）上市后研究[10, 11]，目标值无不适当电击率为 91.6%（相当于不适当电击率为 8.4%），S-ICD 预计可达到 94.6%，α=0.05（单侧），检验效能 90%，计算样本量。考虑到临床研究中因各种原因，最终入选的

样本量与计算的样本量可能会有差异，对实例 1 和实例 3，除了计算样本量，还考察实际入选例数在计算出的样本量±5 范围内的实际成功率与检验效能。

本研究的编程及统计分析均使用 R 语言（v3.6.2）和 Rstudio 平台（v1.2.5033），使用的 R 语言包有 proportion、pwr、TrialSize 及 gsDesign。

结果

1 样本量

不同方法计算的样本量比较见图 1，规律并不十分明确，但大体趋势为在π_0<0.5 时，正态近似法的样本量最小，其次为反正弦法，接着通用法与确切经典法较为接近，最后为确切保守法。在π_0>0.5 时，通用法样本量最小，其次反正弦法与确切经典法较为接近，接着正态近似法的样本量逐渐超过确切保守法。

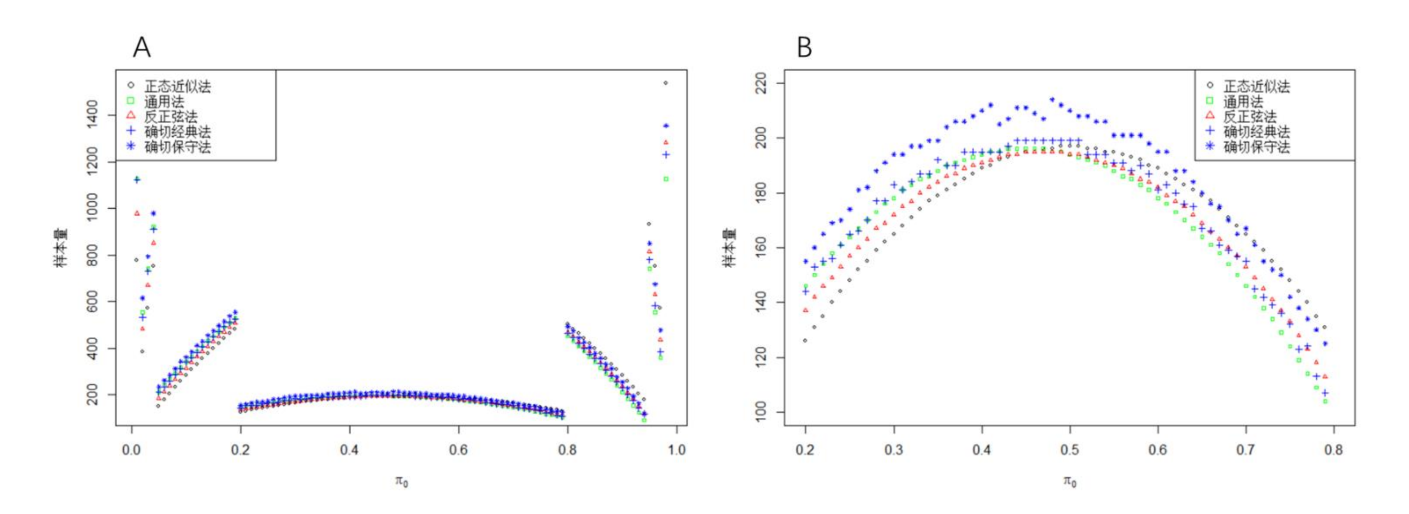


图 1 不同方法计算的样本量比较

横坐标为π_0，从 0.01 至 0.98，按 0.01 递增。π_1>π_0，具体设定见正文。A 为整体视图，B 为 A 的中段（$\pi_0 \in [0.2, 0.8)$）的局部放大。

2 实际成功率

不同样本量计算方法下的实际成功率比较见图 2，可见不同方法所需的实际成功率相差不大，且均略小于π_1。

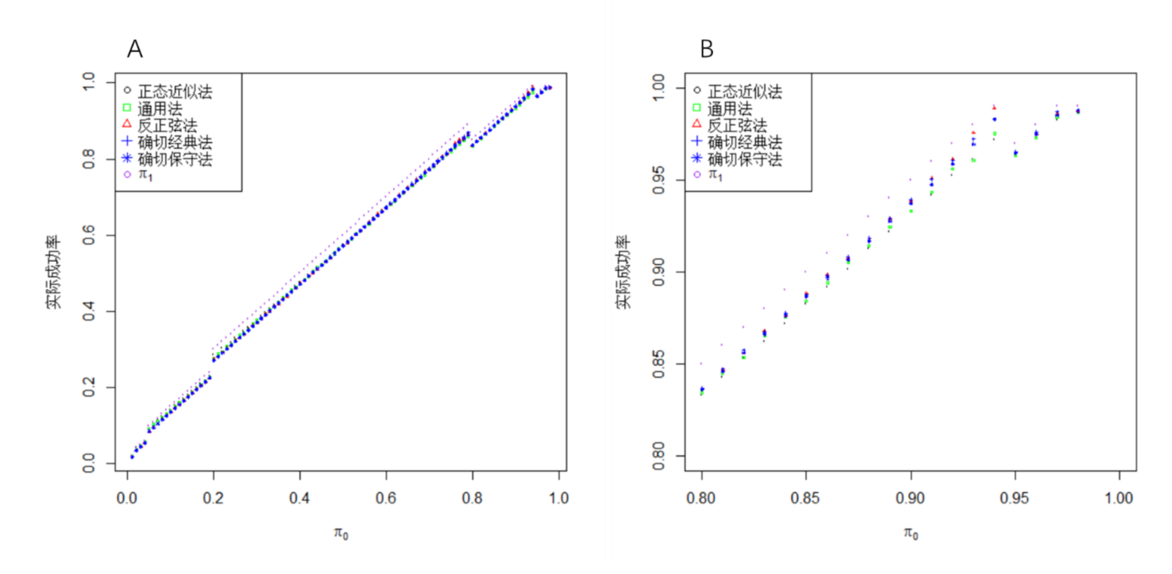


图 2 不同方法的实际成功率比较

横坐标为π_0，从 0.01 至 0.98，按 0.01 递增。$\pi_1 > \pi_0$，具体设定见正文。A 为整体视图，B 为 A 的右侧（$\pi_0 \in$ [0.8 0.98]）的局部放大。

3 检验效能

不同样本量计算方法下的检验效能比较见图 3。显然，不同方法的检验效能差异较大，正态近似法和通用法在π_0<0.5 时检验效能不足，而在π_0>0.5 时检验效能过度，且在π_0趋向极端时，这种趋势急剧增加。反正弦法的检验效能基本在 0.8 附近，但较确切经典法显得更为离散，特别是π_0趋向 1 时。确切保守法几乎能确保检验效能在 0.8 以上，但在π_0>0.8 时，也存在检验效能明显增加的趋势。

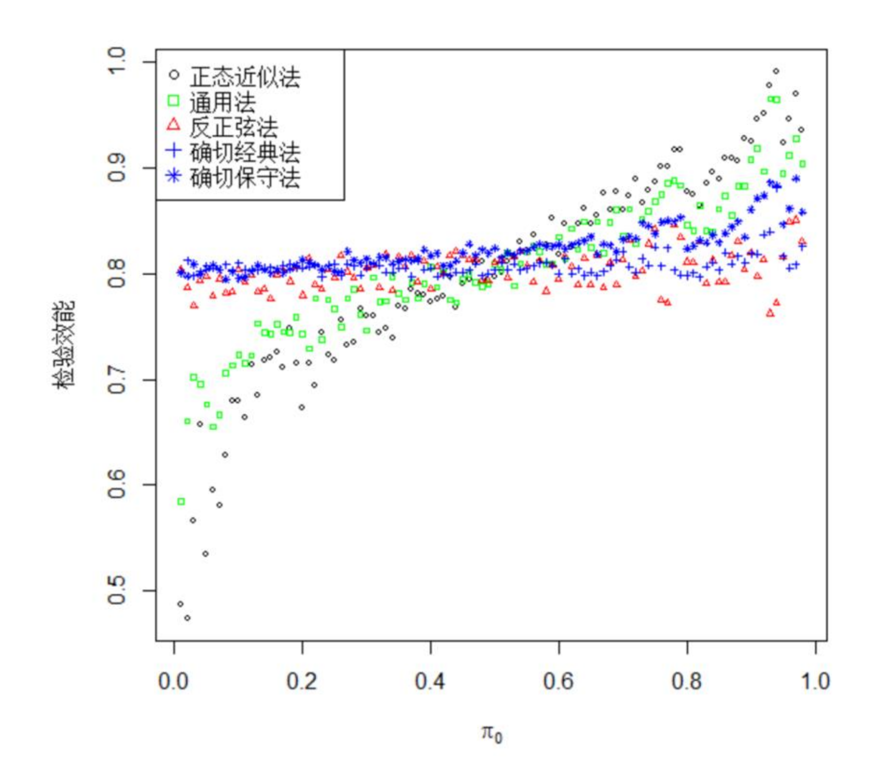


图 3 不同方法的检验效能的比较

横坐标为π_0，从 0.01 至 0.98，按 0.01 递增。π_1>π_0，具体设定见正文。

4 实例分析

不同方法计算的 4 个实例的样本量见表 1，不同方法的样本量有一定的差异，但均与原始文献[7, 9, 10]相对应方法的结果相同（原始文献采用的方法分别为：实例 1 为通用法，实例 2 给出了确切经典法和确切保守法的结果，实例 4 为确切保守法）。对实例 1 和实例 3，样本量±5 范围内的比较结果见表 2，基本符合图 1、图 2 和图 3 显示的趋势，样本量和检验效能的差异较为明显，表现为在π_0<0.5 时，正态近似法和通用法得到样本量相对较小，但损失了检验效能；π_0>0.5 时，正态近似法和通用法样本量相对较大，检验效能也比预设值高。不同方法的实际成功率的差异不大，但较大的样本量可以凭借略为较小的实际成功率通过检验。另外，虽然实际成功率的差异总体不大，但在不同方法间仍然存在一些细小的差异，比如对于实例 1，当样本量为 69 时，正态近似法只需成功 55 例（成功率 79.1%）即可拒绝无效假设，而确切保守法需要成功 57 例（成功率 82.6%）。

表 1 实例分析不同方法计算样本量的结果

实例	π_0	π_1	α(单侧)	Β	正态近似法	通用法	反正弦法	确切经典法	确切保守法
1	0.7	0.85	0.025	0.2	74	64	60	61	70
2	0.07	0.03	0.025	0.2	320	259	224	240	277
3	0.1	0.2	0.025	0.2	71	86	98	94	111
4	0.916	0.946	0.05	0.1	733	619	605	614	668

表 2 不同方法计算样本量的检验效能的比较

π_0/π_1	正态近似法				通用法				反正弦法				确切经典法				确切保守法			
	n	r	P	Power	n	r	P	Power	n	r	P	Power	n	r	P	Power	n	r	P	Power
	69	55	0.7971	0.9162	59	48	0.8136	0.8320	55	45	0.8182	0.8076	56	47	0.8393	0.6731	65	54	0.8308	0.7327
	70	56	0.8000	0.9016	60	49	0.8167	0.8163	56	46	0.8214	0.7954	57	47	0.8246	0.7649	66	54	0.8182	0.8234
	71	57	0.8028	0.8965	61	49	0.8033	0.8909	57	47	0.8246	0.7681	58	48	0.8276	0.7515	67	55	0.8209	0.8020
	72	58	0.8056	0.8890	62	50	0.8065	0.8713	58	47	0.8103	0.8467	59	49	0.8305	0.7251	68	56	0.8235	0.7876
实例 1	73	58	0.7945	0.9286	63	51	0.8095	0.8611	59	48	0.8136	0.8343	60	50	0.8333	0.7215	69	57	0.8261	0.7735
π_0=0.7	74	59	0.7973	0.9179	64	52	0.8125	0.8470	60	49	0.8167	0.8194	61	50	0.8197	0.7978	70	57	0.8143	0.8402
π_1=0.85	75	60	0.8000	0.9176	65	52	0.8000	0.8978	61	50	0.8197	0.7964	62	51	0.8226	0.7845	71	58	0.8169	0.8300
	76	61	0.8026	0.8990	66	53	0.8030	0.8898	62	51	0.8226	0.7870	63	52	0.8254	0.7731	72	59	0.8194	0.8183
	77	61	0.7922	0.9368	67	54	0.8060	0.8742	63	51	0.8095	0.8559	64	53	0.8281	0.7543	73	60	0.8219	0.8015
	78	62	0.7949	0.9323	68	55	0.8088	0.8633	64	52	0.8125	0.8517	65	54	0.8308	0.7361	74	60	0.8108	0.8588
	79	63	0.7975	0.9228	69	55	0.7971	0.9129	65	53	0.8154	0.8319	66	54	0.8182	0.8235	75	61	0.8133	0.8544
	66	13	0.1970	0.5682	81	15	0.1852	0.6707	93	16	0.1720	0.7869	89	16	0.1798	0.7262	106	18	0.1698	0.8139
	67	14	0.2090	0.4757	82	16	0.1951	0.5928	94	16	0.1702	0.8028	90	16	0.1778	0.7497	107	18	0.1682	0.8240
	68	14	0.2059	0.4929	83	16	0.1928	0.6047	95	16	0.1684	0.8134	91	16	0.1758	0.7454	108	18	0.1667	0.8432
	69	14	0.2029	0.5214	84	16	0.1905	0.6323	96	17	0.1771	0.7523	92	16	0.1739	0.7788	109	18	0.1651	0.8482
实例 3	70	14	0.2000	0.5482	85	16	0.1882	0.6448	97	17	0.1753	0.7650	93	16	0.1720	0.7823	110	19	0.1727	0.8017
π_0=0.1	71	14	0.1972	0.5724	86	16	0.1860	0.6713	98	17	0.1735	0.7838	94	16	0.1702	0.7997	111	19	0.1712	0.8032
π_1=0.2	72	14	0.1944	0.5875	87	16	0.1839	0.6876	99	17	0.1717	0.7915	95	17	0.1789	0.7392	112	19	0.1696	0.8238
	73	14	0.1918	0.6243	88	16	0.1818	0.6913	100	17	0.1700	0.8074	96	17	0.1771	0.7502	113	19	0.1681	0.8339
	74	15	0.2027	0.5208	89	17	0.1910	0.6332	101	17	0.1683	0.8149	97	17	0.1753	0.7620	114	19	0.1667	0.8436
	75	15	0.2000	0.5409	90	17	0.1889	0.6532	102	17	0.1667	0.8314	98	17	0.1735	0.7801	115	19	0.1652	0.8567
	76	15	0.1974	0.5727	91	17	0.1868	0.6598	103	17	0.1650	0.8447	99	17	0.1717	0.7933	116	19	0.1638	0.8578

讨论

本文对单个率比较的样本量计算的 5 种不同的方法作了较为系统的研究，5 种方法在π_0及π_1不接近 0 或 1 时表现较为相似，但π_0逐渐接近 0 时，正态近似法和通用法得到的样本量相对较小，并逐渐损失了检验效能；π_0逐渐接近 1 时，正态近似法和通用法的样本量相对较大，检验效能也比预设值逐渐增高。从检验效能来看，反正弦法的结果与确切经典法接近而显得更为离散，而确切保守法几乎能保证预设的检验效能，但在π_0>0.5 时，确切保守法所需样本量比确切经典法逐渐增加。不同方法对实际成功率的要求总体相似，但存在细小差别。

目前临床研究中对于这 5 种不同方法的选用，尚没有明确的适用条件。本研究显示通用法的整体表现优于正态近似法，但这两种方法本质上都是基于正态近似的原理，当π_0趋于极端时，检验效能有较大的变化。有教科书[3]建议当率偏向两侧时（π_0<0.3 或π_0>0.7）使用反正弦法，但从图 3 来看，这个建议并不理想，π_0介于 0.4 和 0.6 之间时检验效能才可保持在 0.8 左右，但这又势必明显限制了这两种方法的临床应用。也有作者[6, 12, 13]认为 nπ及 n(1-π)>5 时，可以考虑正态近似的方法，但是在率较小时，计算的样本量会相应的较大，比较容易满足 nπ及 n(1-π)>5 的条件（如实例 4），故而这个建议也不理想。

不同方法的比较鲜有研究。有作者比较了通用法和确切经典法的表现，认为π_1>0.85 时，确切经典法所得的样本量略低于通用法，且检验效能也低于通用法[14]，本文的结果基本与之一致，并且显示通用法检验效能的提高是以样本量增加为代价的，而此时的检验效能已明显大于预设值了。该研究设定的 π_0从 0.7 开始，因此失去了考察π_0<0.5 的机会，而本文显示π_0<0.5 的表现与>0.5 的表现几乎是相反的。值得一提的是，在 5 种样本量计算的方法中，除了反正弦法，其他 4 种方法中，π_0和π_1的是不可互换的，比如目标值 0.1 和预计值 0.2 的样本量，与目标值 0.2 和预计值 0.1 的样本量是不一样的。

本研究的局限在于：1）研究主要为计算机模拟分析，未进行深入的理论讨论，并且模拟的情形相对有限。但本文的模拟基本覆盖了临床常见的一些情形，并且精选了 4 个实例作了进一步的分析验证。2）临床研究中选择样本量的考虑还有其他许多重要因素，比如目标值的确定、对受试者脱漏的估计、缺失数据的处理等。本研究无法对众多因素一一考量，在对实例 1 和 3 的分析中可以看出，根据确切法计算的样本量，实际入选例数不宜轻易减少，否则检验效能达不到预设值。

综上，单个率比较的样本量计算方法的选择较为复杂，从样本量计算本身来看，对检验效能要求比较高时（如产品的上市前研究），宜优选确切经典法和确切保守法，其次可考虑反正弦法，而通用法和正态近似法在率偏向两侧时，样本量会过大或过小，应具体权衡。

参考文献

[1] 于明坤，明扬，夏如玉，等. 国际目标值法临床研究的文献和方法学特征分析[J]. 中国循证医学杂志, 2019,19(11):1308-1316.
[2] 孙振球，徐勇勇，主编. 医学统计学[M]. 4版. 北京：人民卫生出版社, 2014.
[3] 颜虹，徐勇勇，主编. 医学统计学[M]. 3版. 北京：人民卫生出版社, 2015.
[4] 李卫，赵耐青. 单组目标值临床试验的统计学考虑[J]. 中国卫生统计, 2017,34(03):505-508.
[5] 国家药品监督管理局医疗器械技术审评中心. 医疗器械临床试验设计指导原则[S]. 2018.
[6] 吕德良，李雪迎，朱赛楠，等. 目标值法在医疗器械非随机对照临床试验中的应用[J]. 中国卫生统计, 2009,26(03):258-260.
[7] 刘江美，陈平雁. 单样本率确切概率检验的样本量与检验效能非单调变化关系的研究[J]. 中国卫生统计, 2012,29(02):164-167.
[8] 曾治宇，林娜，张明东，等. 单样本率比较(单组目标值法)的样本量计算及其简便实现[J].

中国卫生统计, 2018,35(02):313-314.
[9] 国家药品监督管理局医疗器械技术审评中心. 人工耳蜗植入系统临床试验指导原则[S]. 2017.
[10] Gold M R, Knops R, Burke M C, et al. The Design of the Understanding Outcomes with the S-ICD in Primary Prevention Patients with Low EF Study (UNTOUCHED)[J]. Pacing Clin Electrophysiol, 2017,40(1):1-8.
[11] Boersma L V, El-Chami M F, Bongiorni M G, et al. Understanding Outcomes with the EMBLEM S-ICD in Primary Prevention Patients with Low EF Study (UNTOUCHED): Clinical characteristics and perioperative results[J]. Heart Rhythm, 2019,16(11):1636-1644.
[12] 成琪, 刘玉秀, 陈林, 等. 单组临床试验目标值法的精确样本含量估计及统计推断[J]. 中国临床药理学与治疗学, 2011,16(05):517-522.
[13] 肖林海, 赵耐青. 单样本率精确概率检验的样本量估计方法及在Stata中的实现[J]. 中国卫生统计, 2014,31(04):711-714.
[14] 唐欣然, 黄耀华, 王杨, 等. 单组目标值试验样本量计算方法的比较研究[J]. 中华疾病控制杂志, 2013,17(11):993-996.